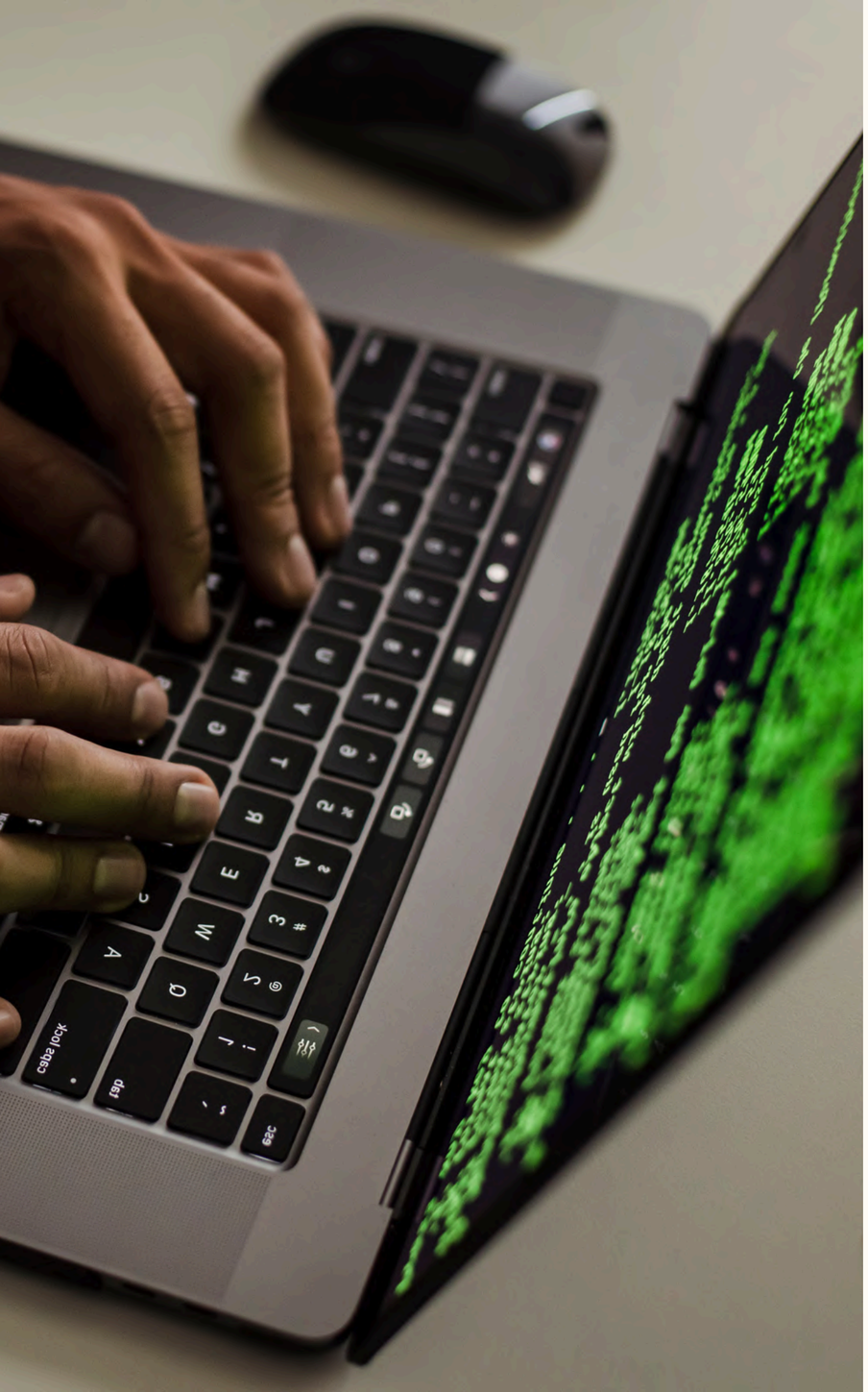
Fraud Detection in Credit Card using Machine Learning

Certified Spesific Independent Study - Data Science & AI

Seftico Frig Injek B
2501984846

Muhammad Fadzli Maula
2502009860

Overview



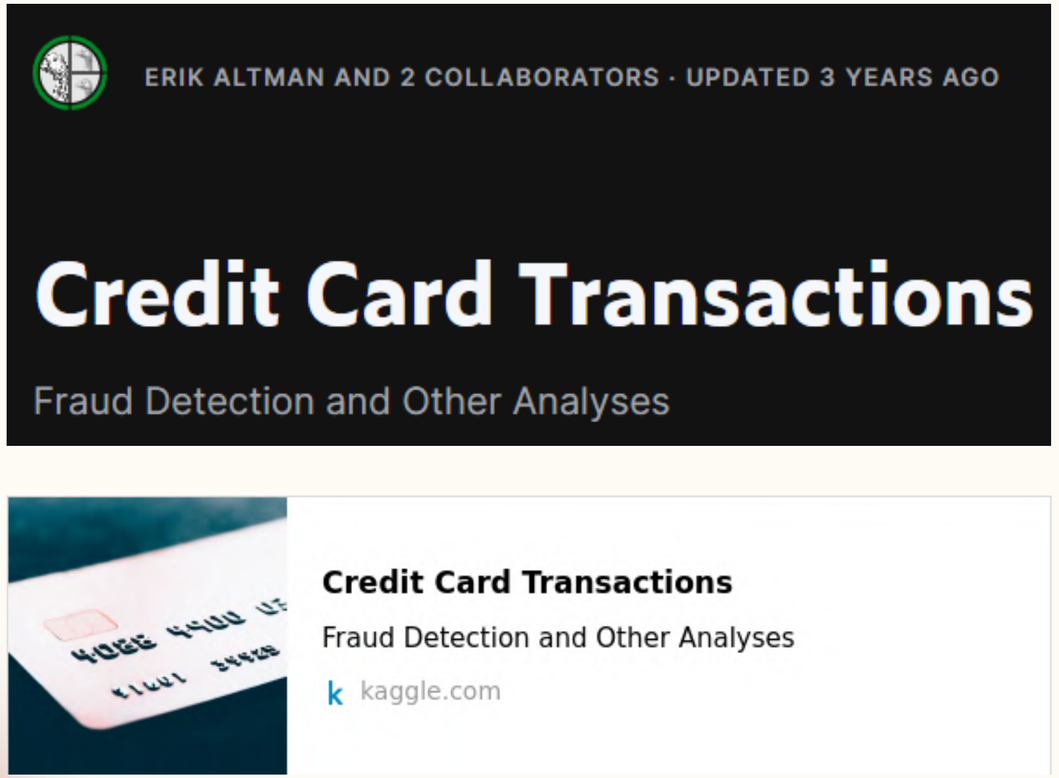
Back to Overview

Introduction

Fraudulent activities in credit card transactions pose a significant threat to financial institutions and customers. The ability to detect fraudulent transactions promptly can save substantial financial losses and enhance security. Our study aims to evaluate the effectiveness of different machine learning models in identifying fraudulent transactions. Specifically, we focus on comparing the performance of Logistic Regression, XGBoost, and Gradient Boosting classifiers.

Dataset Overview

For our analysis, we utilized a publicly available dataset from Kaggle, which includes credit card transactions over a period. This data has more than 20 million transactions generated from a multi-agent virtual world simulation performed by International Business Machines. The dataset was created combining three data sources; users, card and credit card information with the transaction record data.



Back to Overview

Data Processing

Processing the data is a crucial step in preparing it for model training. We handled missing values by using appropriate imputation techniques and scaled the numerical features to standardize the range. Additionally, we encoded categorical variables using one-hot encoding. One significant challenge we faced was class imbalance, as fraudulent transactions are rare compared to legitimate ones. To address this, we employed downsampling techniques to balance the dataset.

Back to Overview

Data Processing

In this section, we will clean and create features within the three different files

User Information

- Create user id "User"
- Remove the dollar sign before "Yearly Income - Person", "Total Debt", "Per Capita Income - Zipcode"
- Create the following features:
  1. Retired: Yes if current age > retirement age
  2. $Person\ Location\ Income\ ratio = \frac{(Yearly\ Income - Person)}{(User\ Location\ Income + 0.01)}$
  3. $Person\ Income\ toDebt = \frac{(Yearly\ Income - Person)}{(Total\ Debt + 0.01)}$
  4. $Location\ Income\ toDebt = \frac{(User\ Location\ Income)}{(Total\ Debt + 0.01)}$
- These variables are created because they can possibly representing the income level and financial pressure of users, which can be associated with fraudulent transaction

Card Information

- Create index "User_Card"
- Join with user information

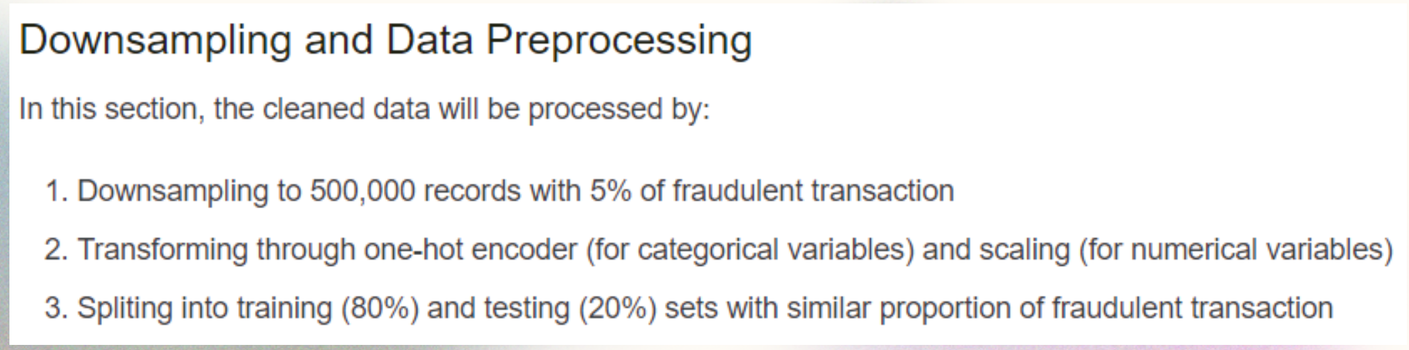
Transaction Data

- Join with card and user information
- Remove the dollar sign before "Amount" and remove data with negative transaction amount
- Remove columns such as errors, merchant name, these columns are either too many missing values or hard to interpret
- Create Day of the week
- Create Time of the day (when the transaction occur, separating a day into 8 periods)
  - Midnight (23-2), Early Morning (2-5), Morning (5-8), Late Morning (8-11), Noon (11-14), Afternoon (14-17), Evening (17-20), Late Night (20-23)
- Create The last digit of tranaction amount

Back to Overview

Downsampling and Data Preprocessing

Downsampling and data preprocessing steps are crucial for preparing the clean data for machine learning algorithms.



By carefully applying these techniques, we can ensure that our models are trained on high-quality data, leading to improved performance, reduced training time, and enhanced interpretability.

Back to Overview

Back to Overview

Benefits of Downsampling and Data Preprocessing

- **Improved model performance:** By addressing data imbalances and preparing the data in a format suitable for machine learning algorithms, these steps can significantly enhance the accuracy and generalizability of the model.
- **Reduced training time**: Downsampling the dataset can shorten the training process, making it more efficient and computationally feasible.
- **Enhanced model interpretability**: Data transformation techniques can make the model's decision-making process more transparent and easier to understand.

Back to Overview

Modeling Techniques

We implemented three machine learning models:
Logistic Regression, XGBoost, and Gradient Boosting Classifier.
Logistic Regression is a linear model commonly used for binary classification tasks.
XGBoost is an efficient and scalable implementation of gradient boosting, known for its high performance in classification problems and,
Gradient Boosting Classifier is an ensemble learning technique that builds models sequentially to improve prediction accuracy.

Model Evaluation

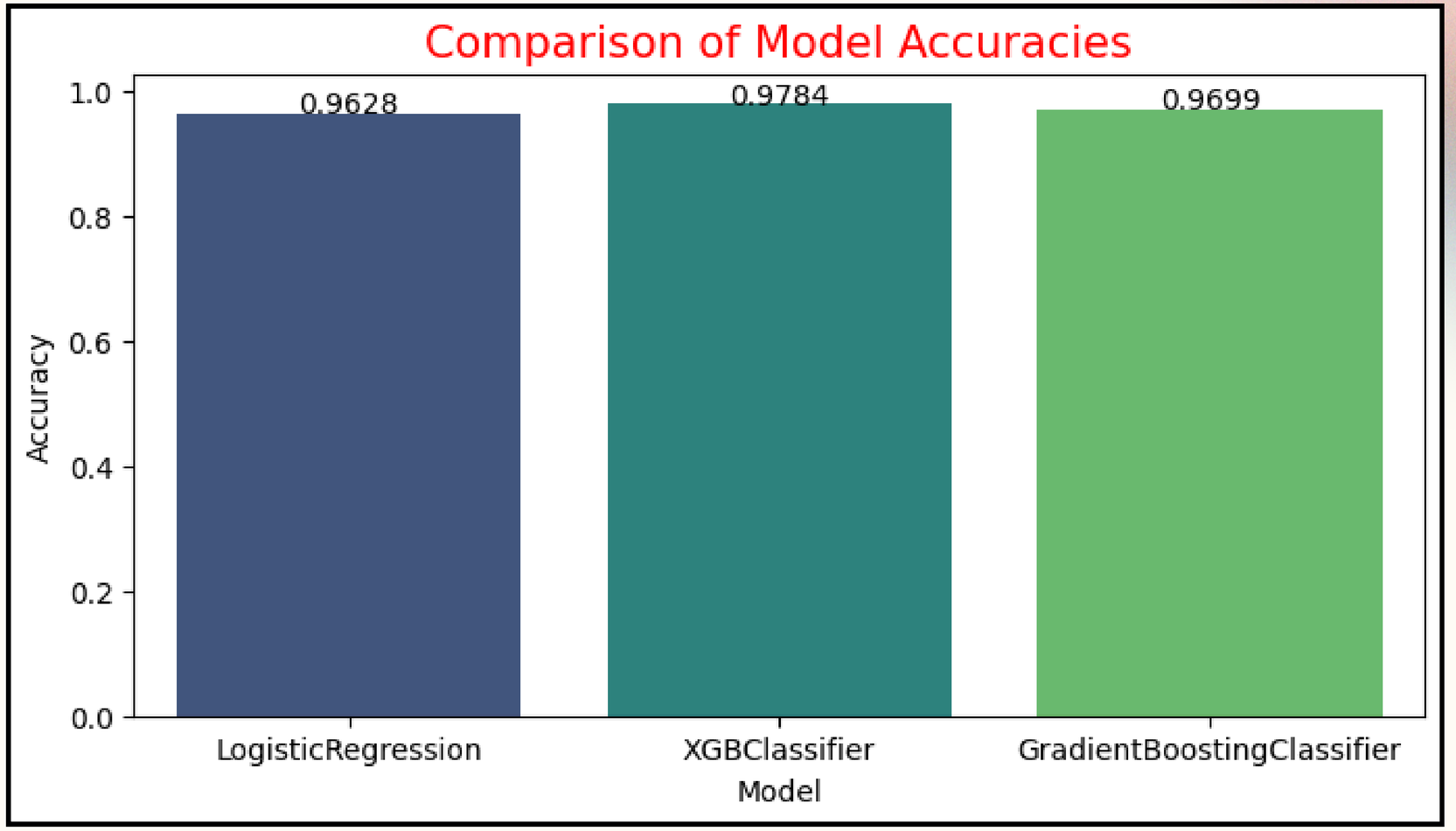
To evaluate the performance of our models, we used several metrics:

- Accuracy: The ratio of correctly predicted instances out of the total instances.
- Precision: The ratio of true positive instances out of all instances predicted as positive.
- Recall: The ratio of true positive instances out of all actual positive instances.
- F1 Score: The harmonic mean of precision and recall, providing a single measure of performance.
- ROC AUC: The area under the Receiver Operating Characteristic curve, indicating the model's ability to distinguish between classes.

Back to Overview

Comparison of Model Accuracies

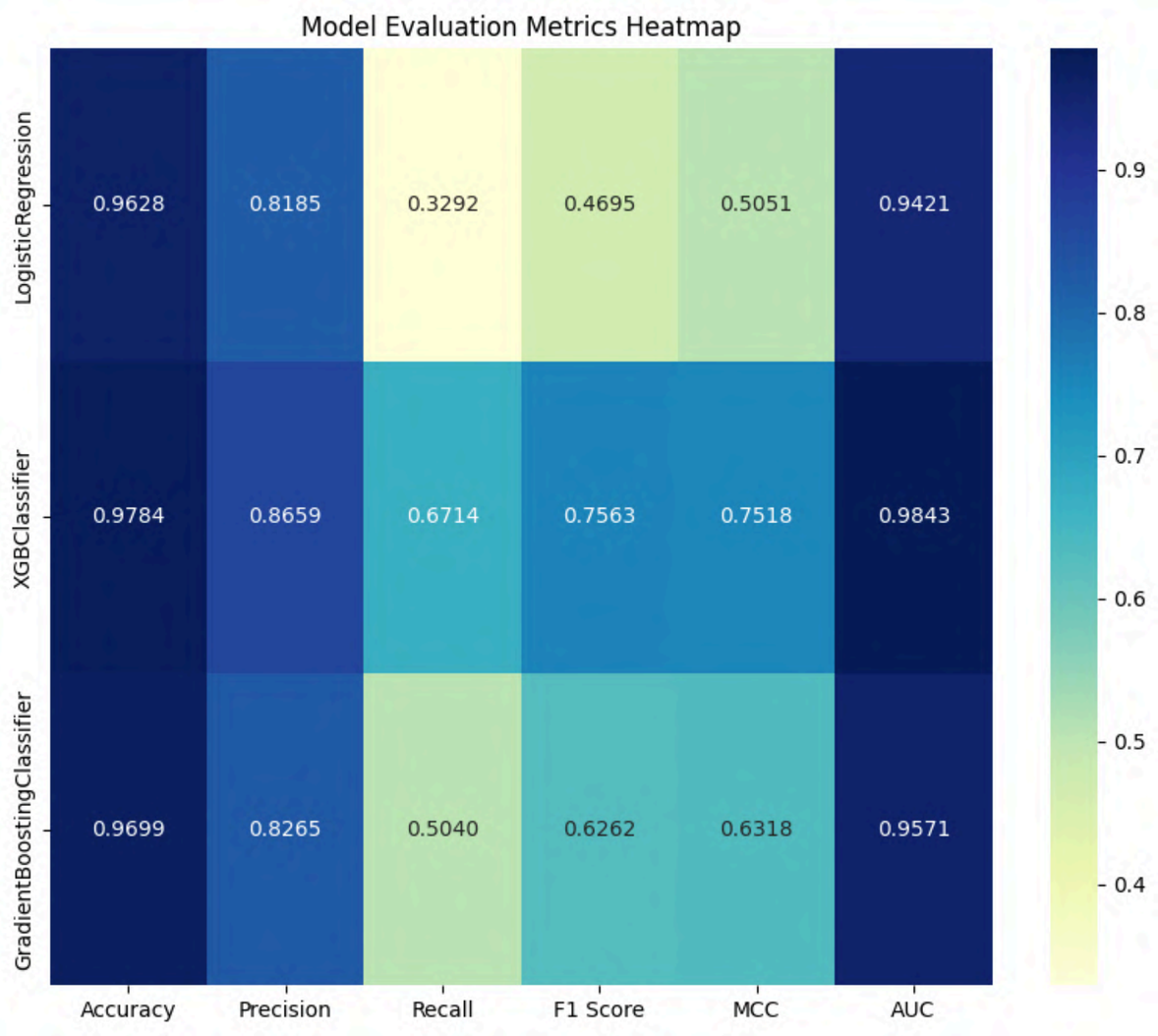
Back to Overview



This visual representation makes it easier to compare the performance of each model at a glance. From the plot, we can observe that certain models, such as XGBoost, achieved higher accuracy compared to others. These insights are crucial as they guide us in selecting the most effective model for fraud detection.

Result

Back to Overview



We can quickly identify the strengths and weaknesses of each model relative to the others. For instance, the XGBoost model consistently exhibits high scores across all metrics, while the Logistic Regression and Gradient Boosting Classifier model shows lower performance in certain areas.

Discussion

Back to Overview

Each model has its strengths and weaknesses. Logistic Regression, being a simpler model, is easier to implement and interpret but may not capture complex patterns as effectively as ensemble methods. XGBoost and Gradient Boosting, while more complex, offer superior performance and are better suited for handling intricate relationships in the data. The insights from our study highlight the importance of choosing the right model for fraud detection tasks, balancing complexity, and performance.

Conclusion

Back to Overview

In conclusion, our study demonstrates that ensemble methods, particularly XGBoost, provide superior performance in detecting credit card fraud compared to Logistic Regression and Gradient Boosting Classifier. Future work could explore further improvements, such as optimizing hyperparameters, exploring different sampling techniques, and incorporating additional features. Our findings underscore the potential of machine learning models in enhancing fraud detection systems and improving financial security.

Thank You

Back to Overview

For a detailed view of the entire project, including code, data processing, and model implementations, please visit my GitHub repository at https://github.com/SefticoFIB/Kaggle-Project